ACOUSTIC-TO-ARTICULATORY INVERSION OF SPEECH: A REVIEW

Asterios Toutios, Konstantinos Margaritis
Parallel and Distributed Processing Laboratory
Department of Applied Informatics
University of Macedonia
156 Egnatia Str., P.O. Box 1591, 54006, Thessaloniki, Greece
{toutios, kmarg}@uom.gr
http://macedonia.uom.gr/~{toutios, kmarg}

Abstract : In this article, we review a specific speech processing research area called acoustic-to-articulatory inversion of speech, or simply speech inversion, which has attracted many researchers and scientists during the last 35 years. The underlying problem refers to the mapping from the acoustic space, which is well-defined since it consists of acoustic signals, to the articulatory space. The latter is somewhat vague, since it may manifest itself in a variety of ways, though it is always assumed that it bears some kind of connection to the human speech production system and the relative position of the articulators - the parts of the human vocal tract. We consider three major classes of models for the articulatory space, be it theoretical models, classic but of little practical interest, medical scanning models, where articulatory data are derived directly from the human subject by means of a specialized device such as an electromagnetic articulograph or an electropalatograph, and, finally, linguistics-derived models, which relate a given phoneme to a vector of phonetic features. A number of approaches have been proposed in the quest for a solution to the speech inversion problem such as codebook approaches, neural network approaches, constrained optimization approaches, analytical approaches as well as stochastic modelling and statistical inference methods such as mixture density networks or Kalman filtering. The recovery of the articulatory space from the speech acoustic signal could have a variety of applications such as building visual aids for hearing impaired people or as a means of study in phonetics and phonology. And most of all, the additional articulatory information made available may improve the performance of current speech recognition systems, and especially in cases such as with noisy, spontaneous or pathological speech. This possibility is demonstrated in several recent papers, where the articulatory information is embedded in speech recognition systems by various means, such as Bayesian Networks or factorial Hidden Markov Models.

Keywords: Acoustic-to-Articulatory Inversion, Speech Recognition, Machine Learning

1. INTRODUCTION

Human speech is, from a mechanical viewpoint, the effect of the airflow from the lungs to the acoustic environment, through the human vocal tract. The variability of the sounds that can be produced, is caused by the level and type of constriction to this airflow imposed by the various parts of the vocal tract, such as the lips, the tongue or the vocal cords, to name a few. These parts of the human vocal tract are called the articulators. Obviously, their positioning plays a critical role in human speech production. We can refer to the positioning of the body of the articulators with the term articulatory state.

On the other side, and still in this mechanical viewpoint, the obvious manifestation of human speech is the acoustic signal. This one we can measure, analyze, process and after all hear.

What is of concern here is the relationship between the articulatory state and the acoustic signal, which we can view as a mapping of two spaces onto each other. These spaces are the articulatory space and the acoustic space.

Actually, if the articulatory state is known, then the corresponding acoustic signal can be easily derived. This is not the case with the inverse problem. The recovery of the articulatory state given the acoustic signal is considered a difficult and ill-posed problem which is puzzling researchers for over three decades now. One of the reasons for this difficulty is the "one-to-many" nature of the acoustic-to-articulatory inversion problem: a given articulatory state has always only one acoustic realization but an acoustic signal can be the outcome of more than one articulatory states. Also, the problem is highly non-linear: two somewhat similar articulatory states may give rise to totally different acoustic signals. These facts come to an extreme in the case of the ventriloquist, where the articulators seem to be static, while a plethora of sounds is being heard.

The motivation behind the ongoing research despite the inherent difficulties seems to arise from the possible applications of a successful solution. Maybe the most interesting is the use of the additional articulatory information derived from such a solution to improve the performance of current speech recognition systems, especially in cases such as with noisy, spontaneous or pathological speech. Other possible applications include speech synthesis, building visual aids for teaching hearing impaired people how to speak and as a means of study in phonetics and phonology.

An important expected outcome of the acoustic-to-articulatory inversion is the modelling of coarticulation – the way the acoustic manifestation of a particular phoneme is dependent of its context. Classic speech recognition approaches deal with this problem mainly by considering biphones or triphones instead of phonemes as classes to be modelled, thus increasing by large the number of these classes, and needing far more training examples.

The rest of this paper is organized as follows: Chapter 2 introduces the ways the articulatory space may be modelled in order to perform the inversion. In chapter 3 we review some of the current proposals for the solution of the acoustic-to-articulatory inversion, or acoustic-to-articulatory mapping, or speech inversion problem. Chapter 4 further explores the relationship between acoustic-to-articulatory inversion and speech recognition. Conclusions are drawn in chapter 5, together with a brief insight into our

future research work.

2.MANIFESTATIONS OF THE ARTICULATORY SPACE

The articulatory space may manifest itself in a variety of ways. Various models may be called in order to describe it. We may view three broad classes of such models. We will call them “theoretical models”, “medical scanning models” and “linguistics-derived” models.

Several theoretical models have been proposed in order to describe the human speech production process, such as Maeda's model [1] or the lossless tube model [2]. Such models were the ones of choice for early works on acoustic-to-articulatory inversion. They are still used in numerous works.

A number of techniques have been developed in order to acquire the articulatory state directly from the human subject by means of some specialized medical scanning device. In x-ray cineradiography [3] the vocal tract of the subject is x-ray filmed during speech production. This particular technique is no longer used because of the danger of radiation exposure, however a lot of old x-ray films have been preserved. In electromagnetic misdagittal articulography (EMMA) or electromagnetic articulography (EMA) [4], sensor coils are attached to the human subject, on places such as the lips, the tongue body or the tongue tip. Then the human subject wears a special helmet that produces an alternating magnetic field. The position of the articulators can thus be recorded. In electropalatography (EPG) [5] the patterns of contact between the tongue and the palate during speech are determined. This technique utilises an artificial palate with 62 silver electrodes embedded in its tongue-facing surface. Finally, in laryngography [6], two electrodes positioned on the throat record the contact variations between the vocal folds of the speaker. The expected results of all these measurements are trajectories of the movement of the articulators – articulatory trajectories – that vary slowly in time.

Obviously, the derivation of such data is a difficult and quite expensive task. However, a number of databases have recently been made available, opening new roads for research in the acoustic-to-articulatory inversion field. These include the MOCHA database [7], with EMA, EPG and laryngography data, the EUR-ACCOR [8] database, with EPG, laryngography and pneumotachography (measurements of nasal and oral airflow velocity), and the X-ray Wisconsin database [9], with EMA-like data. All of these databases include of course, the corresponding acoustic data.

A totally different way to describe the articulatory space is to use knowledge of linguistics, and particularly phonetics. Then, each phoneme of the spoken language is related to a vector of features that describe in a somewhat abstract sense the articulatory state. These features can be either multivalued or binary. Multivalued features often describe the articulatory state in terms of the place and the manner of articulation. An example of such a set of features can be found in [10]. On the other hand, binary features describe the articulatory state as the presence or absence of a specific phonetic quality. The justification of those features is based on the coronal work found in [11]. A third kind of features, the Government Phonology [12] primes have also been used in the same sense.

For deriving from the speech signal the features described in the latter paragraph

often the term “Detection of Phonological Features” instead of “Acoustic-to-Articulatory Inversion” is being used, as they may have a functional, as opposed to a strictly articulatory, meaning. We prefer to have a unified view, believing that there is a great degree of overlap between the two problems.

3.APPROACHES TOWARDS A SOLUTION

One of the first things to consider when building a speech recognition system is the choice of features by which the acoustic signal will be represented, in other words the choice of a suitable front-end parameterization. The same stands for acoustic-to-articulatory inversion. In most such systems, the acoustic features of choice are the Mel Frequency Cepstral Coefficients (MFCCs) [13]. The MFCCs are robust, contain much information about the vocal tract configuration regardless the source of excitation, and can be used to represent all classes of speech sounds. Other features like the Perceptual Linear Predictive (PLP) [14] coefficients may also be used. In [15] an interesting set of acoustic parameters is being presented and their association to specific phonetic features is thoroughly investigated. The implementations however presented here are typically using the MFCCs.

Numerous acoustic-to-articulatory inversion methods use codebook lookup procedures combined with optimization procedures in order to perform the inversion. The articulatory space is quantized and the corresponding acoustic features are synthesized to form a codebook of acoustic/articulatory vector pairs. The quality of the expected result - the articulatory trajectories - is highly dependent on the initial solutions given by the codebook. Thus, it is important that the codebook gives a good coverage of the articulatory space. In [16] the codebook is represented in the form of a hierarchy of hypercubes. Each hypercube represents a region of the articulatory space in which the articulatory-to-acoustic mapping is linear. For each acoustic entry the whole codebook is searched for the relative articulatory parameters to be retrieved.

Another class of approaches to the acoustic-to-articulatory inversion problem is based on the use of neural networks. The parameters of some neural networks are trained to get a nonlinear continuous mapping between the articulatory parameters and the acoustic features. These approaches are most useful when the articulatory space is represented by means of abstract linguistics-derived parameters. In [17] recurrent neural networks are used to perform feature detection on three phonological feature systems, be it binary features, multi-valued features and government phonology primes. The networks perform well, with the average accuracy for a single feature ranging from 86% to 93%. In [10] a set of multilayer perceptrons is used in order to map between MFCC parameters and a set of multi-valued articulatory features.

A constrained optimization approach for estimating the articulatory state from the speech signal is presented in [18]. The scheme used concatenates a gradient search, which is accelerated by using an algorithm inspired by the Fletcher-Reeves method, a classical nonlinear optimization approach, and a linear successive approximation which assures convergence near the optimum articulatory vector. Constraints are imposed on the articulatory parameters to avoid physiologically impossible vocal tract configurations.

An example of an analytic approach to the acoustic-to-articulatory inversion problem

is presented in [19], where a variational calculus method and Maeda's articulatory model are used. The method includes inherent coarticulation constraints in the definition of an energy function to be minimized analytically. [20] is an example of a Linear-predictive based approach. We should also mention at this point, as a part of the analytic approaches, the work by the same author found in [21] where the quest is for a set of acoustic parameters that incorporate articulatory constraints.

Perhaps the most up-to-date and promising class of solutions is the one that is based on stochastic modelling and statistical inference methods. In [22] a mixture density network is called upon to perform the acoustic-to-articulatory inversion while EMA data from the MOCHA database are used. The investigation there showed that the mixture density network is very well suited to delivering the required functionality for performing the inversion mapping. In [23] EMA data are again used but this time the method of choice is dynamical system modelling (Kalman filtering). The speech signal is parameterized by means of linear predictive coding (LPC) analysis [24]. One of the conclusions of this work is that the underlying physical mechanism of speech production is sufficiently linear not to require non-linear models; however, the acoustic observations do not have a linear relationship to the articulatory parameters. In [25] a non-linear filtering approach is taken. This work outlines a stochastic framework for adapting an artificial model to real speech from acoustic measurements alone, using the Expectation Maximization (EM) algorithm [26] and showing that solution of the problem in a maximum-likelihood sense relies on solving an associated state-estimation problem to gather statistics from the measurement data. In [27] the EM algorithm is used again, with the E-step accomplished by the Iterated Extended Kalman filtering [28] and smoothing, to estimate the articulatory model parameters. EMA data are used and the method is tested only on vowel tokens. EPG articulatory data from the ACCOR database and PLP parameterization of the acoustic signal are used in [29], where a latent variable approach to the acoustic-to-articulatory mapping is presented. In latent variable modelling, the combined acoustic and articulatory data are assumed to have been generated by an underlying low-dimensional process. A parametric probabilistic model is estimated and mappings are derived from the respective conditional distributions.

Before closing this chapter, we have to point out the work found in [30], where methods for applying phonetic and phonological constraints to provide unique solutions to the acoustic-to-articulatory inversion problem are reviewed and discussed upon.

4.ARTICULATORY INFORMATION FOR SPEECH RECOGNITION

Current speech recognition systems [30, 31] typically use Hidden Markov Models (HMMs), Neural Networks or hybrid schemes in order to map between the acoustic speech signal and the corresponding words or phonemes. A language model is used to retrieve the a priori probabilities of the appearance of these language units. Apart from this language model, the only input source of such systems is the acoustic signal, parameterized in some way. These systems achieve satisfactory results in the case of normal, structured and noise-free speech. This is not the case with noisy, spontaneous, or pathological speech.

It is widely accepted that systems based on this classic approach have reached a plateau in terms of performance. And, since they do not completely satisfy us, new

approaches need to be discussed. One of them uses articulatory information in order to enhance recognition. This information cannot be readily available for everyday applications and has to be retrieved by means of acoustic-to-articulatory inversion. In the following, we present some recent works that explore the use of articulatory information in the context of speech recognition.

In [10] multi-valued abstract articulatory features extracted from the speech signal by means of a set of multilayer perceptrons are used as a source of information for recognition of clear, reverberant and noisy speech. Three different input sources for the recognition task are considered: acoustic features alone, articulatory features alone, and both of them simultaneously. The system corresponding to this last input source derives as a combination of the previous two systems by means of a product rule. The results indicate that using the articulatory features alone doesn't improve much recognition using acoustic features: the results are somewhat similar. However, the combined system exhibits such an improvement, especially in the noisy speech case: as a matter of fact the improvement increases as the speech-to-noise ratio gets lower.

EMA data and linear dynamic modelling is used in [33]. Both real and simulated articulatory data are considered for a phone recognition task. The conclusion is actually the same as in the previous case: the use of combined acoustic and articulatory information improves recognition performance.

The authors of [34] introduce a type of HMM in which each state represents an articulatory configuration. The state transition matrix is governed by dynamic constraints on articulator motion. They call this scheme a Hidden-Articulator Markov Model, or HAMM. The model itself doesn't produce better word recognition results compared to an acoustics-based standard HMM, however a combination of the two systems does. It is suggested that the articulatory system makes in general different mistakes than the acoustic one; a fact that is actually beneficial for the recognition task.

Another study [35], building upon the groundwork done in [36], investigates the use of dynamic Bayesian networks (DBNs) for incorporating articulatory data with acoustic data in automatic speech recognition. During training, the articulatory data, which are derived from the X-ray Wisconsin database, are introduced as variables to the DBN, which is expected, during testing, to be able to infer the distribution of the articulatory positions given the observed acoustics, thus accomplishing the acoustic-to-articulatory inversion task as a sub-product of the recognition task itself.

Finally, the authors of [36] attempt to use knowledge of abstract binary articulatory features in the context of recognizing dysarthric speech.

5.CONCLUSION AND FUTURE WORK

The recovery of the articulatory space from the acoustic signal poses an interesting problem which has attracted, and keeps attracting the interest of researchers worldwide. The problem is not a trivial one and the approaches towards its solution range within a wide spectrum of methods and techniques, mostly from the artificial intelligence field of study. The recent availability of articulatory databases gives an extra boost to the relative research.

Perhaps the most interesting field of application for the acoustic-to-articulatory inversion is speech recognition. It has indeed been proven that knowledge of the

articulatory state can enhance the performance of speech recognition systems; an improvement that is actually needed given the current state of such systems. The role of the inversion in this context is to provide such data. Work is still in an early stage; a fully functional speech recognition system that uses articulatory information is yet to be developed.

In this paper, we have reviewed some of the most recent approaches to speech inversion. For a discussion of earlier ones, the interested reader may refer to [38]. Surely, our discussion here is by far a non-exhaustive one since the field of acoustic-to-articulatory inversion is quite large, with more than a hundred of published works so far. We have just outlined some basic concepts, trying to explain what speech inversion is all about.

In the future, we are planning to look further into the acoustic-to-articulatory mapping problem, which we view as an interesting and demanding machine learning problem, beginning with recreating some of the experiments described in the works mentioned above. Our main concern is with stochastic modelling, statistical inference and neural network methods. Our long-term goal is to perform successfully the inversion for some of the English-speaking articulatory databases described above and then use the extracted information in the context of a Greek speech recognition system, exploiting the global nature and language-independence of the acoustic-to-articulatory mapping. We are also working with abstract linguistics-derived features in the same sense.

References

1.S. Maeda. Un Modèle Articulatoire de la Langue avec des Componsantes Linéaires. Actes 10èmes Journées d' Etude sur la Parole, Grenoble, pp. 152-162,1979.

2.J. D. Markel and Jr. A. H. Gray, Linear Prediction of Speech, Springer Verlag, Berlin, 1976.

3.K. G. Munhall, E. Vatikiotis-Bateson and Y. Tokhura, X-Ray Film Database for Speech Research, Journal of the Acoustical Society of America, 98, pp. 1222-1224, 1995.

4.J. Ryalls, Introduction to Speech Science: From Basic Theories to Clinical Applications, Allyn & Bacon, 2000.

5.W. J. Hardcastle, The Use of Electropalatography in Phonetic Research, Phonetica, 25, pp. 197-215, 1972.

6.S. Winstanley and H. Wright, Vocal Fold Contact Area Patterns in Normal Speakers: An Investigation using the Electrolaryngograph Interface System, British Journal of Disorders of Communication, 26, pp. 25-39, 1991.

7.A. A. Wrench and W. J. Hardcastle, A Multichannel Articulatory Database and its Application for Automatic Speech Recognition. In Proceedings 5th Seminar of Speech Production, pp. 305-308, Kloster Seeon, Bavaria, 2000.

8.A. Marchal, W. Hardcastle, P. Hoole, E. Farnetani, A. Ni Chasaide, O. Schmidbauer, I. Galiana-Ronda, O. Engstrand, and D. Recasens, EUR-ACCOR: The Design of a Multichannel Database, Actes du XIIème Congres International des Science Phonétiques, Aix-en-Provence, 5, pp. 422-425, 1991.

9.J. R. Westbury, X-Ray Microbeam Speech Production Database User's Handbook. University of Wisconsin, Madison, 1994.

10.K. Kirchoff, Robust Speech Recognition Using Articulatory Information, PhD. Thesis, University of Bielefeld, Germany, 1999.

11.N. Chomsky and M. Halle. The Sound Pattern of English, MIT Press, 1968.

12.J. Harris, English Sound Structure, Blackwell, 1994.

13.S. B. Davis and P. Mermelstein, Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences, IEEE Transactions Audio, Speech and Signal Processing, 28, 357-366, 1980.

14.H. Hermansky, Perceptual Linear Predictive (PLP) Analysis of Speech, Journal of the Acoustical Society of America, vol. 87, no 4, pp. 1738-1752, 1990.

15. A.V Hansen, Acoustic Parameters Optimised for Recognition of Phonetic Features, In Proceedings Eurospeech-97, pp. 397-400, Rhodes, Greece, 1997.

16. S. Ouni and Y. Laprie, Improving Acoustic-to-Articulatory Inversion by using Hypercube Codebooks, In Proceedings ICSLP2000, Bejing, China, 2000.

17. S. King and P. Taylor, Detection of Phonological Features in Continuous Speech using Neural Networks, Computer Speech and Language, 14(4), pp. 333-353, 2000.

18. P. P. L. Prado, E. H. Shiva and D. G. Childers, Optimization of Acoustic-to-Articulatory Mapping, In Proceedings ICASSP'92, vol. 2, pp. 33-36, 1992.

19. Y. Laprie and B. Mathieu, A Variational Approach for Estimating Vocal Tract Shapes from the Speech Signal, In Proceedings ICASSP'98, pp. 929-932, 1998.

20. S. Krstulovic, LPC Modelling with Speech Production Constraints, In Proceedings 5th Speeech Production Seminar, 2000.

21. S. Krstulovic, Speech Analysis with Production Constrains, Ph.D. Thesis, Ecole Polytechnique Fédérale de Lausanne, 2001.

22. K. Ricmond, Mixture Density Networks, Human Articulatory Data and Acoustic-to-Articulatory Inversion of Continuous Speech, In Proceedings Workshop on Innovation in Speech Processing WISP'2001, 2001.

23. S. King and A. Wrench, Dynamical System Modelling of Articulator Movement, In Proceedings ICPhS'99, San Francisco, USA, 1999.

24. L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, 1978.

25. G. Ramsay, A Non-linear Filtering Approach to Stochastic Training of the Articulatory-Acoustic Mapping Using the EM Algorithm, In Proceedings ICSLP'96, 1996.

26. J. Bilmes, A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Technical Report, University of Berkeley, 1997.

27. S. Dusan and L. Deng, Recovering Vocal Tract Shapes from MFCC Parameters, In Proceedings ICSLP'98, Sydney, Australia, 2000.

28. A. M. Jazwinsky, Stochastic Processes and Filtering Theory, Academic, New York, 1970.

29.M.A. Carreira-Perpinan and S. Renals, A Latent Variable Modelling Approach to the Acoustic-to-Articulatory Mapping Problem, In Proceedings ICPhS'99, San Francisco, USA, pp. 2013-2016, 1999.

30.S. Dusan, Methods for Integrating Phonetic and Phonological Knowledge in Speech Inversion, Proceedings of the International Conferense on Speech, Signal and Image Processing, Malta, 2001.

31.L. R. Rabiner, A Tutorial on Hidden Markov Model and Selected Applications in Speech Recognition, Proceedings of the IEEE, 77(2), pp. 257-286, 1989.

32.R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen and V. Zue (eds.), Survey of the State of the Art in Human Language Technology, Center for Spoken Language Understanding, Carnegie Melon University, Pittsburgh, USA, 1996.

33.J. Frankel and S. King, Speech Recogntion in the Articulatory Domain: Investigating an Alternative to Acoustic HMMs, In Proceedings Workshop for Innovations in Speech Processing, 2001.

34.M. Richardson, J. Bilmes and C. Diorio, Hidden-Articulator Markov Models for Speech Recognition, ISCA ITRW Conference on Automatic Speech Recognition, Paris, France, 2000.

35.T. A. Stephenson, H. Bourlard, S. Bengio and A. C. Morris, Automatic Speech Recognition using Dynamic Bayesian Networks with Both Acoustic and Articulatory Variables, In Proceedings ICSLP2000, Bejing, China, 2000.

36.G. G Zweig, Speech Recognition with Dynamic Bayesian Networks, Ph.D. Thesis, University of California, Berkeley, 1998.

37.N. Sawhney and S. Wheeler, Using Phonological Context for Improved Recognition of Dysarthric Speech, Project Report 6345, MIT Media Lab, 1999.

38.J. Scrhroeter and M. M. Sondhi, Techniques for Estimating Vocal-tract Shaoes from the Speech Signal, IEEE Transactions Speech and Audio Processing, 2(1), pp. 133-150, 1994.